**Developing a Policy QA System Using Retrieval-Augmented Generation (RAG) Framework Literature Review**

**1. Overview**

RAG (Retrieval-Augmented Generation) is a method to improve the capabilities of large language models (LLMs) by combining knowledge from external databases to enhance model performance. The core lies in the three links of retrieval, generation and enhancement, first retrieving information related to the input sequence from the external knowledge base, and then using LLMs to generate answers based on the retrieved information, and combining external knowledge with the internal knowledge of LLMs to improve the accuracy and credibility of the generated content. RAG technology has gone through a development phase from Naive RAG to Advanced RAG to Modular RAG, continuously optimizing system performance and functionality. There are two main forms of RAG model: RAG-Sequence and RAG-Token, the former always retrieves documents with the same conditions during the generation of sequences, and the latter selects different documents when generating each target marker, both of which marginalize the retrieved documents as latent variables and integrate them into the generation process.

**2. The background and concept of RAG**

The concept of RAG was proposed by the Microsoft Research team in 2021 based on pre-trained language models, aiming to solve the performance bottleneck problem of large-scale pre-trained language models in knowledge-intensive tasks by introducing a retrieval mechanism, enhance the knowledge acquisition ability of the model, and then improve its performance in downstream tasks.

**3. Elaboration of relevant concepts**

KG (Knowledge Graph): A form of structured data that is often used to provide accurate information. In the multi-document Q&A task, nodes represent paragraphs or document structures, and edges represent semantic or lexical similarities between paragraphs, helping large language models establish logical associations between multiple documents.

LLM (Large Language Model): A pre-trained language generation model based on large-scale text data, which is good at generating coherent text content. In the multi-document Q&A task guided by the knowledge graph, the ability to understand and answer complex questions is improved with the help of prompts and prediction mechanisms, combined with the contextual information provided by the knowledge graph.

**4. Working principle**

Retrieval stage: RAG uses the retrieval machine to retrieve documents related to the input sequence from external knowledge bases or text databases by using semantic similarity calculations and other methods to make up for the shortcomings of LLMs in processing queries beyond the scope of training data.

Build Phase:

a. RAG-Sequence: The same retrieved document is marginalized as a latent variable to obtain seq2seq probability p(y|x), the output sequence is generated by top-K approximation and marginalization, and the target sequence is generated by combining the input sequence and the retrieved document.

b. RAG-Token: Different documents can be selected for each target tag and marginalized, so that the generator can integrate multi-document content when generating answers, and combine input sequences and retrieved documents to generate target sequences.

Decoding Phase:

a. RAG-Token: As a standard autoregressive seq2seq generator, it is decoded by standard beam search.

b. RAG-Sequence: Each document needs to be bundled searched and scored, i.e., "thoroughly decoded", and a "fast decode" approach can be used for efficiency, avoiding additional forward passing of assumptions from all candidate set Y.

Training stage: The parameters of the retriever and generator are jointly trained, and the retrieved documents are marginalized into latent variables during the training process, resulting in the distribution of generated text.

Enhancement phase: RAG continuously updates and integrates knowledge from external databases to ensure that the content generated is timely; Components can also be improved in a modular way, such as adding a similarity search module or optimizing a searcher, fusing external and intrinsic LLMs knowledge, improving the accuracy and credibility of generation, reducing the probability of generating incorrect content, and optimizing the performance of knowledge-intensive tasks.

**5. RAG technology application and scenario problem solving**

Used technology:

a. Retrieval: Retrieve relevant document fragments from external knowledge bases through methods such as semantic similarity calculation to obtain the latest and accurate information related to user queries, making up for the shortcomings of LLMs in processing queries beyond the scope of training data.

b. Generation: Combining retrieved external knowledge and user queries, use a generator to generate accurate and informative answers, including two methods: RAG Sequence and RAG Token.

c. Enhancement: Dynamically updating and integrating domain specific information, integrating external and LLMs' internal knowledge, improving the accuracy and credibility of generated content, and reducing the probability of generating incorrect content.

d. Generation flexibility: combines the generation flexibility of "closed volume" (parametric) methods with the performance of "open search" methods.

e. No need for significant span mask pre training: Unlike REALM and T5+SM, RAG can perform well without expensive "significant span mask" pre training.

f. Retrieval initialization: RAG's retrieval is initialized using DPR's retrieval supervisor, which supervises on natural problems and TriviaQA.

g. Document reordering and reader replacement for extraction: By generating answers, RAG can generate correct answers even if they are not in any retrieval documents, with an accuracy rate of 11.8% on the NQ dataset and 0% for the extraction model.

Resolve scenario issues:

a. Dealing with outdated information: Addressing the issue of LLMs' "illusion" when processing data beyond training or when current information is needed, ensuring the timeliness of generated content, and improving the applicability and reliability of real-time and accurate information demand scenarios such as news queries and professional consulting.

b. Knowledge intensive tasks: In scenarios that require complex and current knowledge, such as open domain Q&A and abstract Q&A, RAG can provide more accurate and detailed answers, especially in scenarios where multiple document information needs to be integrated and accurate answers can be generated without explicit answer extraction.

c. Multimodal data processing: Expand the interpretation and processing of various data forms such as images, videos, and codes, and enhance the application capabilities of LLMs in the multimodal field.

d. Marginalized documents: For documents containing answer clues but without clear answers, RAG can effectively marginalize these documents to generate correct answers, which standard extraction methods cannot achieve.

e. No document answer generation: Even if there is no correct answer in the retrieved document, RAG can still generate the correct answer, enhancing the robustness and adaptability of the model.

f. Factual text generation: Compared to pure parametric models such as BART, RAG generates more specific and factual answers, reducing the phenomenon of "illusion" and improving the credibility of content.

**6. Challenges and shortcomings faced by RAG**

Content generation: may generate false or misleading content, or impersonate others on news and social media, as well as automatically generate spam and phishing content, requiring corresponding regulatory and preventive mechanisms.

Knowledge sources: External knowledge sources such as Wikipedia that rely on a large amount of practical knowledge cannot completely avoid bias and errors, which may affect the quality of generated content; And the external databases relied upon may contain outdated information, resulting in answers that are not novel or accurate enough.

Model performance: The inference process is complex and opaque, making it difficult to track and interpret generated decisions, which poses challenges for model debugging, optimization, and user trust.

Insufficient ability to process long texts: Performance may decrease when processing long texts, making it difficult to fully understand and utilize the information in long texts, limiting its application in tasks that require lengthy comprehension.

Multimodality and Evaluation:

a. Multimodal data processing: Although it has expanded to the field of multimodal data processing, further research and improvement are still needed in this area, and existing technologies may not be able to meet the needs of multimodal data processing well.

b. Imperfect evaluation methods: Existing evaluation methods cannot fully measure the actual contribution of RAG, and a more accurate and representative evaluation mechanism is needed to provide strong basis for model improvement and development.

**7. The Evolution History and Latest Technologies of RAG Technology**

Evolution process:

a. Early development: Initially, the RAG model combined pre trained parametric memory (such as seq2seq model) and non parametric memory (such as Wikipedia's dense vector indexing) for language generation tasks, improving generation accuracy and diversity by using retrieved documents as input.

b. Naive RAG：Mainly relying on retrieving and integrating relevant document fragments from external knowledge bases into large language models to enhance their generation capabilities, with a focus on improving pre trained models.

c. Advanced RAG：With the development of large-scale language models, RAG technology has further advanced, not only providing better information for LLMs in the inference stage, but also combining model fine-tuning techniques to better handle complex and knowledge intensive tasks.

d. Modular RAG：In the latest development stage, modular design is used to increase system flexibility and scalability, not only processing text data but also expanding into multimodal fields.

Latest technology:

a. Multimodal Expansion: The application scope of RAG extends to the multimodal field, capable of interpreting and processing various data forms such as images, videos, and code.

b. Improvement of evaluation method: With the expansion of RAG application scope, the evaluation method has been improved to ensure accurate and representative performance evaluation.

c. Technology integration: RAG further leverages technology integration with other AI methods such as fine-tuning and reinforcement learning to gain empowerment.

d. No need for significant span mask pre training: Unlike REALM and T5+SM, RAG can perform well without expensive "significant span mask" pre training.

e. Multi document utilization: RAG effectively utilizes documents containing answer clues but without clear answers to generate correct answers, significantly improving the ability to generate correct answers in multiple knowledge intensive NLP tasks, surpassing the level of pure parametric seq2seq models and task specific retrieval and extraction architecture techniques.